VIVOLAB-UZ Audio Segmentation System for Albayzín Evaluation 2010

Diego Castán, Alfonso Ortega, Carlos Vaquero, Antonio Miguel, Eduardo Lleida

Aragon Institute of Engineering Research (I3A)
University of Zaragoza (UZ)
(dcastan | ortega | cvaquero | amiguel | lleida)@unizar.es

Abstract

This paper presents a method for audio segmentation that separates broadcast news audio files into five acoustic classes for Albayzín Audio Segmentation 2010 [1]. The proposed system makes use of a presegmentation stage based on the Bayesian Information Criterion (BIC), a music/speech classifier based on a combination of GMMs and a binary decision tree, and finally a speech/speech with music/speech with noise classifier based on GMMs.

Index Terms: Audio segmentation, Bayesian Information Criterion, Gaussian Mixture Models, C4.5 Tree

1. Introduction

Segmentation plays an important role in audio processing applications, such as content-based audio retrieval recognition and classification, and audio database management. Audio segmentation is a process that divides an audio file into different classes. Each segment or clip should consist of a single class that is acoustically different from other classes of the audio file. A good segmentation should be able to delimit the boundaries between two classes to group segments into homogeneous classes.

There are many different tasks in audio segmentation. Several methods are focus on speech/music segmentation, speaker recognition or acoustic events detection. Albayzín 2010 Audio Segmentation database is composed of 87 hours of sound (24 files of approximately 4 hours long), will be splitted into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). This evaluation contains five classes:

1. Music (MU): Music is understood in a general sense.
2. Speech (SP): Clean speech in studio from a close microphone.
3. Speech with music in background (SM): Overlapping of speech and music classes or speech with noise in background and music classes.
4. Speech with noise in background (SN): Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
5. Other (OT): This class refers to any type of audio signal (including noises) that doesnât correspond to the other four classes.

This work has been partially supported by the national project TIN2008-06856-C05-04

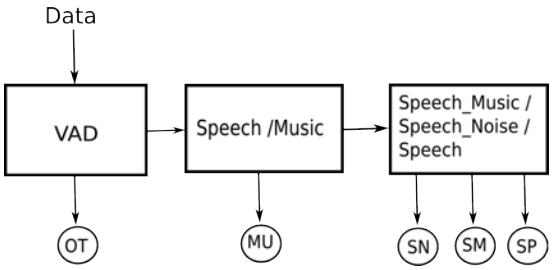

Figure 1: Classification System

The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

The proposed metric is inspired on the NIST metric for speaker diarization. The metric is defined as a relative error averaged over all acoustic classes:

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}$$

The remainder of the paper is organized as follows: section 2 discusses the set of features and masks used in the system, section 3 describes the classification algorithm, section 4 provides an evaluation of the system and is followed by the conclusion in section 5.

2. Features and masks

The features used in the proposed system are 13 MFCCs plus their derivative and second derivative (delta and delta delta).
In order to obtain more discriminative speech-music classification system, the mean, the variance and the skewness of the first MFCC is calculated.

A presegmentation is made by using a sliding window BIC based algorithm to delimit boundaries between heterogeneous segments. At the end, the segmentation system proposed in this work will assign just with one label for each segment obtained by this presegmentation. Long silence segments (with more than 3.5 s) are detected and marked as Others since those segments do not correspond to any of the predefined classes.

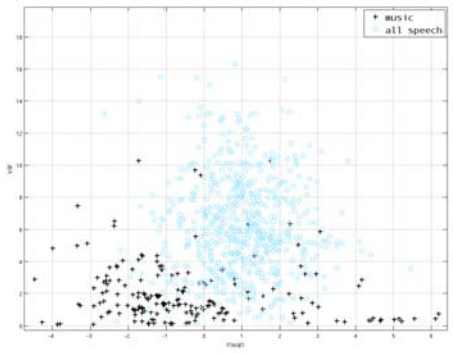

Figure 2: Behavior of the C1-Average VS C1-Variance with speech signal (circles) and music signal (crosses)

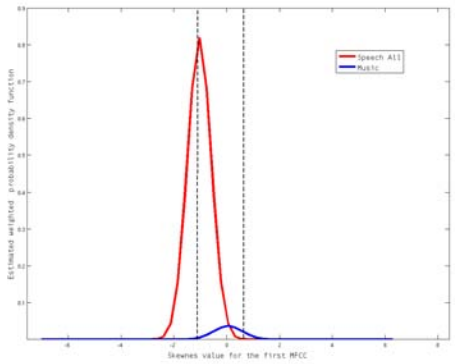

Figure 3: Estimated weighted probability density function of C1-Skewness

3. System Description

The classification of the segments reported by the BIC based segmentation system is made in two steps: a music/speech classification subsystem and a speech/speech with noise/speech with music classification subsystem presented in Fig. 1.

3.1. Music/Speech Classification

This approach is based on two Gaussian Mixture Models (GMM) with 128 components each model. The first model is trained with the music class and the second model is trained with the speech, speech with music and speech with noise classes. The accumulated likelihoods have a penalty based on the a priori knowledge of the distributions of classes within the database.

The GMMs decisions are combined with a binary decision tree trained with the C4.5 algorithm [3]. The tree is trained with the mean, the variance and the skewness calculated every 3 seconds of the first cepstral coefficient that provides information about the distribution of the energy between low and high frequency in a frame. We can see a scatter plot of the C1-mean and the C1-variance in Fig. 2. Also, in Fig. 3 the estimated probability density function of C1-skewness, weighted by the prior probability of speech and music is shown for both classes. It has been evaluated considering constant length non-overlapping windows of 3 seconds. Along with the pdf two lines are plotted dividing the x axis into 3 regions: the one on the right side, contains the points representing the segments that would be correctly classified as music. The one on the left side, contains the points representing the segments that would be correctly classified as speech. In between, the points representing the segments that would need the help of other features to decide if they correspond whether to speech, whether to music.

The combination of both classifiers is made by weighted addition of the GMM likelihood and the tree error prediction of the music class frame by frame.

3.2. Speech, Speech With Music and Speech With Noise Classification

The Speech, Speech with Music and Speech with Noise Classification approach is based on Gaussian Mixture Models (GMM) with 128 components. The first model is trained with the speech class, the second model is trained with the speech with music and the third model is trained with speech with noise class. As in the previous music/speech classifier, these likelihoods have a penalty based on the a priori knowledge of the distributions of classes within the database.

4. Results

The system has been tested over 8 files (around 32 hours of audio material). These results are presented in table 1.

Class	Accuracy
MU	28.14%
SP	51.06%
SM	48.78%
SN	51.51%
Total	44.87%

Table 1: Results on the test files

5. Future lines

Further work must be done to improve the performance of the proposed system along the following lines of research:

- Seek new ways of modelling the temporal behaviour of the detection problem under study. Along this line we think about improving the boundaries estimation substituting the BIC based approach by an HMM based solution.
- Study new ways of representing music and speech in a more dicriminative way. Provided that informal tests seem to show that using statistical moments analysis makes possible to segregate music from speech as shown in Fig. 2 and 3.

- Keep working on the hierarchical structure of our system extending it also for speech classes, and try to explode in a more discriminative way the information extracted from the binary tree classifier in order to fuse it with the information coming from the GMMs.

6. References

[1] Butko, T , "Albayzín Evaluations 2010: Audio Segmentation",

[2] Quatieri T.F., "Discrete-Time Speech Signal Processing", Prentice-Hall, Englewood Cliffs, NJ,USA, 2001.

[3] Quinlan, R., "C4.5: Programs for Machine Learning", in Morgan Kaufmann Publishers Springer Netherlands, 1993.